ANÁLISIS ESTADÍSTICO DEL COMPORTAMIENTO DEL PRIMER ETIQUETADOR CUBANO EN TRES DIFERENTES CORPUS DE LA PRENSA

Leonel Ruiz Miyares
Centro de Lingüística Aplicada
Ministerio de Ciencia, Tecnología y Medio Ambiente, Santiago de Cuba, Cuba
leonel@lingapli.ciges.inf.cu

Larisa Zamora Matamoros
Departamento de Matemáticas
Facultad de Ciencias Naturales y Matemáticas
Universidad de Oriente
Santiago de Cuba, Cuba

Resumen

Cuba ha desarrollado el ETIPROCT (**ETI**quetador y **PRO**cesador de **C**orpus **T**extuales), sistema computacional que procesa grandes corpus textuales.

El ETIPROCT ya se ha aplicado con una alta efectividad a un corpus escrito de estudiantes de secundaria básica, a un corpus textual oral de la comunidad santiaguera de *Los Hoyos* y a un corpus textual escrito de alumnos de primaria del sector rural de la provincia de Santiago de Cuba; sin embargo, este etiquetador no había sido comprobado en un corpus de la prensa escrita.

En este trabajo se realiza un análisis estadístico del etiquetador ETIPROCT a partir del procesamiento en el mismo de tres diferentes corpus de la prensa escrita cubana. Se procesaron varios artículos de los tres periódicos nacionales *Granma*, *Juventud Rebelde* y *Trabajadores*, los cuales trataron sobre las siguientes temáticas periodísticas: educativa, política, social, deportiva, científica, internacional y cultural, lo que permitió obtener una valoración cuantitativa y cualitativa de la efectividad del etiquetador en cada rama de la prensa analizada.

Introducción

El Grupo de Lingüística Computacional del Centro de Lingüística Aplicada de la Delegación Territorial del Ministerio de Ciencia, Tecnología y Medio Ambiente en Santiago de Cuba, ha desarrollado el sistema computacional ETIPROCT (**ETI**quetador y **PRO**cesador de **C**orpus **T**extuales), el cual puede procesar grandes corpus textuales.

El ETIPROCT -etiquetador probabilístico, confeccionado sobre la base de los Modelos Ocultos de Markov (HMM)- ya se ha aplicado con una alta efectividad a un corpus escrito de estudiantes de secundaria básica, a un corpus textual oral de la comunidad santiaguera de *Los Hoyos* y a un corpus textual escrito de alumnos de primaria del sector rural de la provincia de Santiago de Cuba; sin embargo, este tagger no había sido comprobado en un corpus de la prensa escrita.

En este trabajo se realiza un análisis estadístico del etiquetador ETIPROCT a partir del procesamiento en el mismo de tres diferentes corpus de la prensa escrita cubana.

Breve descripción del ETIPROCT

El punto de partida para crear el etiquetador fue la investigación *Léxico Activo Funcional del Escolar Cubano* desarrollada en el Centro de Lingüística Aplicada. Ese trabajo aportó toda la información necesaria para desarrollar las herramientas del futuro tagger.

Sobre la base del material disponible, se decidió que el etiquetador fuera supervisado, probabilístico, utilizara bigramas y aplicara los modelos ocultos de Markov para la desambiguación.

El ETIPROCT trabaja con un conjunto de etiquetas que en nuestro caso asciende a 36 componentes los cuales se describen en la Tabla 1.

Etiqueta	Descripción de la etiqueta
010	Artículo
020	Sustantivo Propio
021	Sustantivo Común Masculino Singular
022	Sustantivo Común Femenino Singular
023	Sustantivo Común Masculino Plural
024	Sustantivo Común Femenino Plural
025	Sustantivo Diminutivo
026	Sustantivo Aumentativo
027	Sustantivo Propio (Nombres Geográficos)
028	Sustantivo Propio (Juegos Infantiles)
030	Adjetivo Calificativo Antepuesto
031	Adjetivo Calificativo Pospuesto
032	Adjetivo Determinativo
033	Adjetivo Diminutivo
034	Adjetivo Aumentativo
040	Pronombre Personal
041	Pronombre Demostrativo
042	Pronombre Posesivo
043	Pronombre Indefinido
044	Pronombre Relativo
045	Pronombre Interrogativo y Exclamativo
046	Variante Pronominal
050	Verbo Modo Indicativo (Tiempos Simples)
051	Verbo Modo Indicativo (Tiempos Compuestos)
052	Verbo Modo Subjuntivo (Tiempos Simples)
053	Verbo Modo Subjuntivo (Tiempos Compuestos)
054	Verbo Modo Imperativo (Tiempos Simples)
055	Perífrasis Verbales
056	Verbo con Enclítico
060	Adverbio
070	Preposición
080	Conjunción
090	Interjección
100	Contracción
110	Lexías Complejas y Fechas
120	Siglas

Tabla 1. Conjunto de etiquetas que utiliza el etiquetador cubano.

El lexicón posee 57 329 palabras que se han añadido a partir del *Léxico Activo Funcional del Escolar Cubano*, del *Diccionario Ortográfico del Español,* desarrollado por el Centro de Lingüística Aplicada de Santiago de Cuba y el Instituto para los Circuitos Electrónicos de Génova, Italia, además de las palabras resultantes de la aplicación del etiquetador en su primera versión a otras investigaciones.

El lexicón contiene también un campo donde se reflejan las características semánticas de las palabras complejas en su significado. No todas las palabras tienen información en ese campo, sólo aquellas como *carata* (botánico), *richter* (sismología), *mira* (militar), lb-12 (gas ecológico), etc. que pueden resultar dudosas para los lingüistas. Este aspecto es de gran utilidad para los lexicólogos en sus estudios de corpus textuales.

A partir de la información obtenida en el léxico escolar, se realizaron los respectivos cálculos estadísticos de los bigramas y de esta forma se confeccionó la matriz de probabilidades de transiciones, según la fórmula:

$$P(t_i/t_{i-1}) \approx \frac{f(t_{i-1},t_i)}{f(t_{i-1})}$$

donde $f(t_{i-1}, t_i)$ es la frecuencia de ocurrencia de la pareja de etiquetas (t_{i-1}, t_i) y $f(t_{i-1})$ es la frecuencia de ocurrencia de la etiqueta t_{i-1}.

Esta matriz juega un importante papel durante la desambiguación de las palabras que posean más de una etiqueta.

En los modelos ocultos de Markov surge la matriz de probabilidades de observación la cual es aquella que calcula la probabilidad de ocurrencia de una palabra dada una etiqueta. Esta matriz se representa según la fórmula:

$$P(w_i/t_i) = \frac{f(w_i, t_i)}{f(t_i)}$$

donde $f(w_i, t_i)$ es la frecuencia de la palabra w_i con la etiqueta t_i y $f(t_i)$ es la cantidad de palabras con la etiqueta t_i.

Para desambiguar las palabras homónimas el etiquetador escoge el mayor resultado de las multiplicaciones de la matriz de transiciones y la matriz de probabilidades de observación para cada caso específico:

$$\max \prod_{i}^{n} P(w_i/t_i)P(t_i/t_{i-1})$$

La muestra

Se procesaron 121 artículos de los tres periódicos nacionales cubanos *Granma* (61 textos), *Juventud Rebelde* (44 textos) y *Trabajadores* (16 textos) publicados entre julio-septiembre de 1999 y mayo-junio del 2000; los cuales fueron recuperados de INTERNET. (Las direcciones electrónicas son: http://granma.co.cu,http://www.jrebelde.cubaweb.cu/ y http://trabajadores.cubaweb.cu/)

Las esferas periodísticas que abordaron los artículos fueron siete y son las siguientes: educativa (10), política (20), social (11), deportiva (26), científica (15), internacional (21) y cultural (18).

El corpus de palabras procesado fue de 45 859 entre los 121 textos analizados y la distribución por esferas periodísticas es como sigue:

	Educativa	política	social	deportiva	científica	internacional	cultural	**Total**
Granma	5	11	6	13	8	11	7	**61**
J. Rebelde	4	7	4	9	5	9	6	**44**
Trabajadores	1	2	1	4	2	1	5	**16**
Total	**10**	**20**	**11**	**26**	**15**	**21**	**18**	**121**

Análisis estadístico de los resultados

La efectividad del ETIPROCT al procesar las 45 859 palabras de los tres periódicos fue de un 97.05%.

En la Fig.1 (ver anexo) se muestra la distribución de la efectividad del ETIPROCT por esferas de la prensa escrita cubana.

En esta figura podemos observar que el etiquetador, a pesar de tener una alta efectividad en general para la prensa cubana, es más efectivo en las esferas deportiva (97.38%), educativa (97.38%) y política (97.36%); encontrándose dicha efectividad por debajo del 97% en las esferas social (96.96%), científica (96.91%), cultural (96.71%) e internacional (96.31%).

Una de las causas que inciden en el por ciento de efectividad obtenido en la esfera científica, es la aparición de palabras en otro idioma, principalmente en inglés, lo que dificulta el procesamiento de los textos. Ejemplo de lo anterior se observa en el artículo del periódico *Juventud Rebelde* publicado el 30 de septiembre de 1999, titulado *Descubren clave para tratamiento de dolores*, el cual en una de sus partes dice *'Los resultados fueron publicados por la revista especializada* ***Proceedings of the National Academy of Sciences'.*** En este caso el ETIPROCT realiza una falsa etiquetación, como es de suponer.

En la Fig.2 se muestra la distribución de la frecuencia de los errores más significativos cometidos por el tagger en la prensa cubana.

Es necesario esclarecer en esta figura que las secuencias representadas por ejemplo de la siguiente manera 050/021 significan que el ETIPROCT codificó la palabra con la categoría 050 (verbo modo indicativo (tiempos simples)) en lugar de la categoría gramatical 021 (sustantivo común masculino singular).

En la Fig.2 resalta que el error con mayor frecuencia incurrido por el etiquetador fue el 031/030 (108 veces), es decir, codificó como adjetivo calificativo pospuesto a palabras que eran adjetivos calificativos antepuestos. Sin embargo, si analizamos con profundidad este tipo de error nos damos cuenta que el etiquetador reconoció la palabra como adjetivo y no como otra categoría gramatical, lo que puede ocurrir en algunos casos, por lo que consideramos esta equivocación como medio error.

Algunas palabras que pueden clasificarse como adjetivos u otras categorías gramaticales extraídas del corpus de la prensa escrita y que fueron codificadas como adjetivos pospuestos por el etiquetador en lugar de adjetivos antepuestos son las siguientes: *mayor (*puede ser: *adjetivo o sustantivo) '...donde cobrará **mayor** importancia...', largo (*puede ser: *adjetivo o sustantivo) '...y a más **largo** plazo...', alto (*puede ser: *adjetivo, interjección o adverbio),'...tienen un **alto** costo...'; mortal (*puede ser: *adjetivo o sustantivo) '...esta **mortal** enfermedad...'; vivo (*puede ser: *adjetivo o verbo) '...se realiza el **vivo** homenaje...'; mejor (*puede ser: *adjetivo o adverbio) '...lograr una **mejor** preparación...'; etc.*

La Fig.3 muestra la distribución de frecuencia de los cinco errores más significativos que cometió el tagger en cada esfera periodística. Se puede observar que, además del 031/030, ya analizado, los errores más frecuentes fueron el 021/031 y el 050/031, lo que se debe a un reconocimiento erróneo de las palabras desconocidas

Considerando que la efectividad de un etiquetador es satisfactoria a partir de un 96%, nos propusimos verificar estadísticamente la eficacia del ETIPROCT en cada una de las esferas periodísticas analizadas y en la prensa en general, para lo cual realizamos la prueba de hipótesis relativa a la proporción, esto es, nos propusimos docimar:

$$H_0: \ p < 0.96 \quad H_A: \ p \geq 0.96 \qquad (1)$$

donde **p** representa la efectividad en términos probabilísticos, H_0 es la hipótesis nula que plantea que la efectividad es inferior al 96% y H_A es la alternativa que se desea verificar y plantea que la efectividad es superior o igual al 96%.

Los resultados de la aplicación de esta prueba se muestran en la Tabla 2:

Esferas	$\hat{p}$	n	$Z_{obs.}$	$Z_{0.99}$
1. Científica	0.9611	6078	3.620396	2.326348
2. Cultural	0.9671	6559	2.934348	2.326348
3. Deportiva	0.9738	10026	7.051432	2.326348
4. Educación	0.9738	4470	4.708334	2.326348
5. Internacional	0.9631	5799	1.204683	2.326348
6. Política	0.9736	8816	6.516420	2.326348
7. Social	0.9696	4111	3.141083	2.326348
Prensa	**0.9705**	**45859**	**11.47456**	**2.326348**

Tabla 2: Resultados de la aplicación de la prueba de hipótesis

En la primera columna de esta tabla se muestran las diferentes esferas analizadas en los tres corpus estudiados, en la segunda las probabilidades estimadas de no cometer errores durante la etiquetación, la tercera columna señala el número de palabras procesadas, la cuarta el valor del Z observado para la dócima (1), esto es,

$$Z_{obs} = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0 q_0}{n}}},$$

donde:

- **n** representa el número de palabras procesadas,
- p_0=0.96, mínimo valor para considerar como bueno a un etiquetador,
- $\hat{p} = \frac{\text{Número de palabras bien etiquetadas}}{n}$,

y la última columna muestra el valor del Z crítico, el cual para un nivel de confiabilidad del 99% es igual a 2.326348.

De las dos últimas columnas de esta tabla podemos observar que para siete de los ocho casos analizados, la dócima es significativa, es decir, que con un nivel de confiabilidad de 99% podemos aceptar que la efectividad del etiquetador es superior al 96% en cada esfera periodística y, por supuesto, en la prensa en general, excepto en la esfera internacional, en la cual la dócima es no significativa si la comparamos con una efectividad del 96% como en los casos anteriores; sin embargo, ella es significativa al compararla con una efectividad del 95% ($Z_{obs.}$ =4.57721)

Este resultado también se puede corroborar en la siguiente tabla:

	N	Media	LI -99,000%	LS +99,000%	Mínimo	Máximo	Varianza
Esferas	7	97.00143	96.43272	97.57014	96.31000	97.38000	0.164714

la cual muestra el número de esferas analizadas (N), por ciento medio de efectividad por esfera (media), límite inferior (LI) y superior (LS) del intervalo confidencial al 99%, los valores mínimo y máximo y la varianza.

De la tabla podemos concluir que la efectividad por esfera oscilará entre un 96.43% y un 97.57% con un nivel de confiabilidad del 99%.

Antes de concluir este trabajo, sería interesante comparar el ETIPROCT con otros etiquetadores ya existentes. Por ejemplo, la versión en español del etiquetador Xerox Tagger [Sánchez F. y Nieto A., 1995] posee 55 etiquetas, emplea los HMM y tiene una efectividad mayor del 97%; el QTAG tagger [Mason O., 1997] es un etiquetador que utiliza sólo principios probabilísticos, utiliza 49 etiquetas y su efectividad es del 97%. La diferencia principal del ETIPROCT con respecto a los etiquetadores señalados estriba en la existencia del aspecto semántico en su lexicón y el amplio espectro de resultados linguo-estadísticos que ofrece, aspecto éste que por razones de espacio no se pudo explicar en este trabajo.

Conclusiones

Con el presente estudio podemos decir que se cierra un ciclo de aplicaciones del primer etiquetador cubano, el ETIPROCT, para demostrar su efectividad en diferentes tipos de corpus textuales. Muestras de textos escritos por escolares urbanos y rurales tanto de primaria como de secundaria básica, textos orales del pueblo y por último de la prensa escrita cubana -esfera diametralmente opuesta a las primeras-, han sido procesadas por el ETIPROCT con resultados satisfactorios lo que demuestra que la etiquetación automática en Cuba va por caminos seguros.

En este trabajo se han mostrado los resultados estadísticos del comportamiento de la aplicación del etiquetador en distintos artículos de los periódicos *Granma*, *Trabajadores* y *Juventud Rebelde*.

A pesar de que los componentes primarios del etiquetador probabilístico ETIPROCT -matriz de transiciones, matriz de probabilidades de observación, etc.- se confeccionaron a partir del estudio del vocabulario del escolar cubano, cuyas características gramaticales y principalmente léxicas presentan algunas diferencias respecto al vocabulario de los adultos, se ha podido comprobar una alta efectividad durante su aplicación al corpus periodístico.

Un elemento que ha propiciado el correcto funcionamiento del tagger es la retroalimentación de su *lexicón* con todas aquellas palabras nuevas incorporadas al mismo luego de ser aplicado a los diferentes corpus y depuradas por los lingüistas, esa medida garantiza que siempre el *lexicón* se actualice y amplíe.

Se pudo comprobar estadísticamente que la efectividad por esferas y en la prensa en general es superior al 96% con un nivel de confiabilidad de un 99%, menos en la esfera internacional, en la cual se pudo comprobar que la efectividad es superior al 95% con un nivel de confiabilidad del 99%.

El análisis estadístico de los errores por esferas permitirá perfeccionar en un futuro el trabajo del ETIPROCT para aumentar aún más su efectividad.

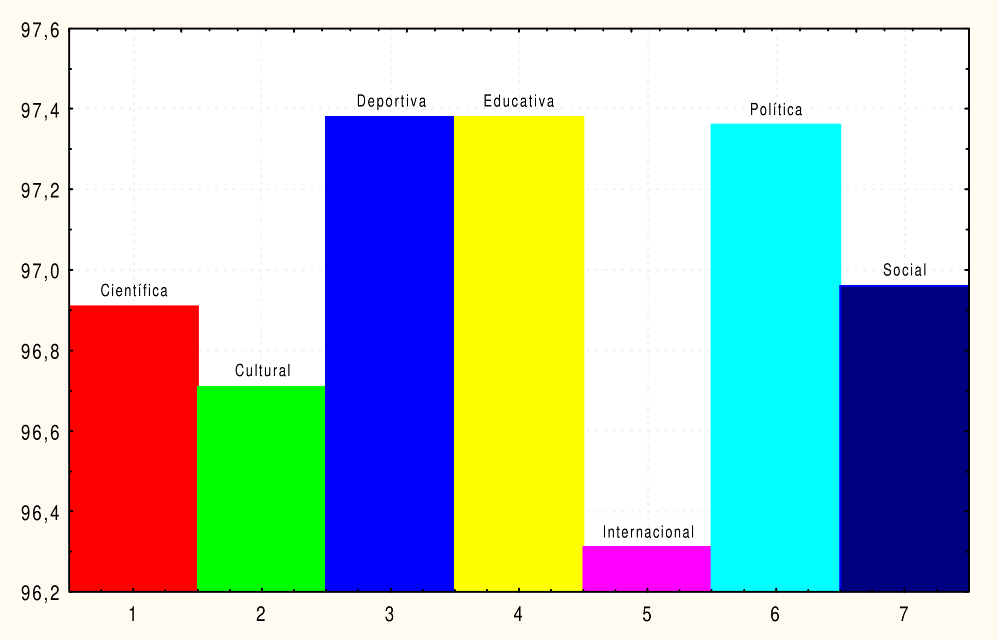

Fig.1 Efectividad del ETIPROCT en las distintas esferas de la prensa escrita.

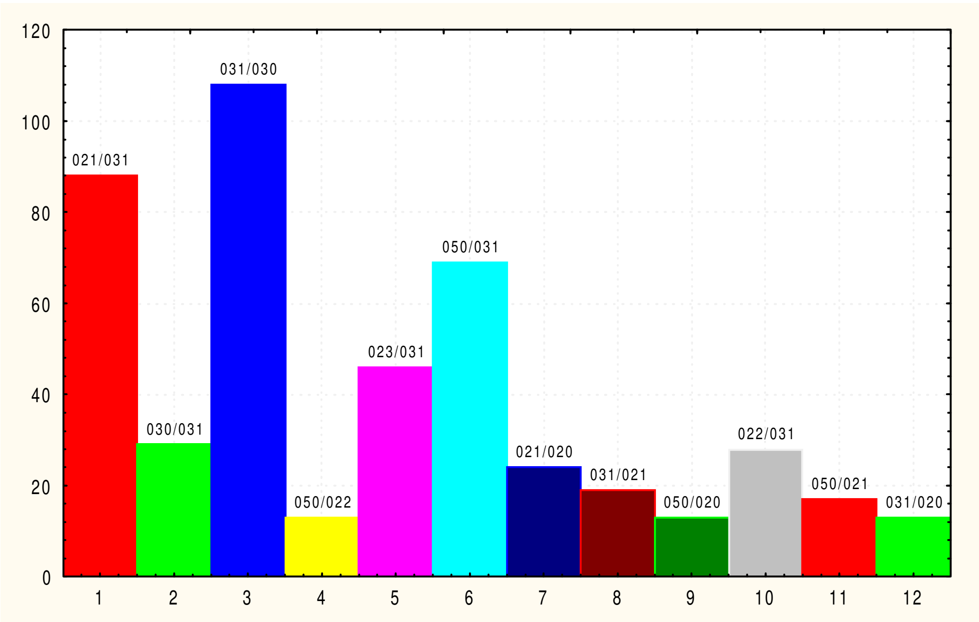

Fig.2 Distribución de la frecuencia de los errores más significativos en la prensa.

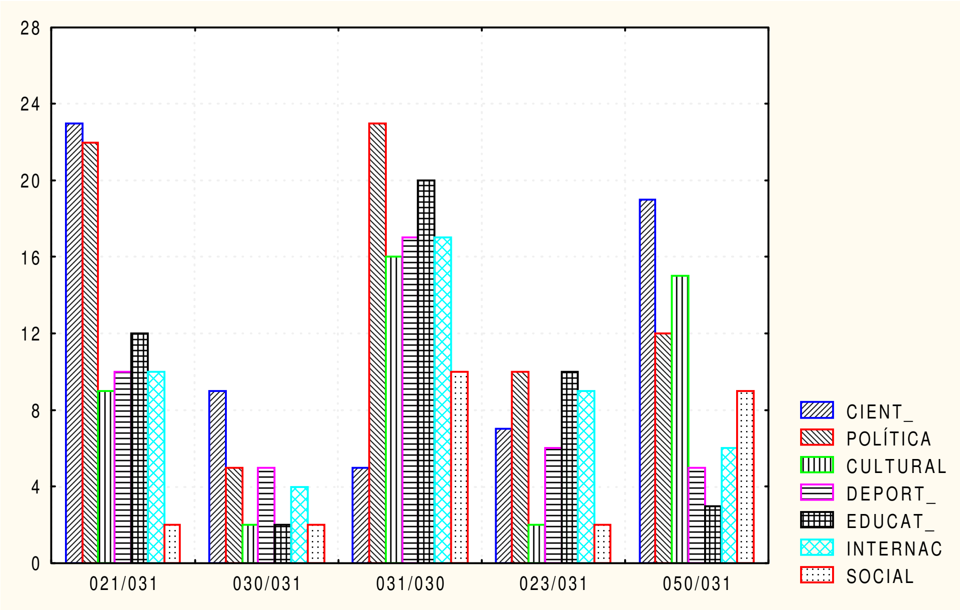

Fig.3 Distribución de la frecuencia de los errores más significativos por esferas.

Bibliografía

Alphen P. (1992). *HMM-based continuous-speech recognition. Systematic evaluation of various system components***.** Doctoral thesis, University of Amsterdam, The Netherlands.

Charniak E. (1993). *Statistical language learning*. Massachusetts Institute of Technology, Massachustts, United States of America.

Mason O. (1997). *QTAG- A portable probabilistic tagger*, 1997. http://www-lg.bham.ac.uk/QTAG/doc/.

Padró Ll. (1997). *A hybrid environment for syntax-semantic tagging*. Tesis de doctorado, Universidad Politécnica de Cataluña, España.

Parzen E. (1972). Procesos estocásticos. Editorial Paraninfo, Madrid, España.

Paulussen H. (1992). *Automatic grammatical tagging: description, comparison and proposal for augmentation*. Universidad de Antwerpen, Wilrijk, Bélgica.

Ruiz L. (1994). *Aplicación de la computación al estudio del vocabulario básico del escolar cubano*. Estudios de Comunicación Social, Editorial Academia, La Habana, páginas 96-105.

Ruiz L. (1997a). *Versión avanzada de un sistema computacional aplicado a una investigación lexicológica*. Estudios de Comunicación Social, Editorial Academia, La Habana, páginas 85-113.

Ruiz L. (1997b). *Development of two probabilistic morphological taggers for Spanish corpus. Evaluation.* Internal Report, University of Twente, Enschede, The Netherlands.

Ruiz L. (1999). *Primeros pasos de la etiquetación automática en Cuba* en las Actas del VI Simposio Internacional de Comunicación Social, Santiago de Cuba, 25-28 de enero de 1999. Ediciones Editorial Oriente, Centro de Lingüística Aplicada y el Consiglio Nazonale delle Riceche, páginas 710-714.

Ruiz L. (2000). *Etiquetación automática en corpus textuales cubanos. Primeros resultados* en las ACTAS del JADT2000, 5tas. Jornadas internacionales de análisis estadístico de corpus textuales, Lausana, Suiza, pp.237-244.

Sánchez F. (1987). *El etiquetado del Corpus de Referencia del Español Actual*

(CREA). Seminario Internacional de Industrias de la Lengua, Soria, España.

Sánchez F. y Nieto A. (1995). *Development of a Spanish version of the Xerox tagger*. http://xxx.lanl.gov/ps/cmp-lg/9505035.